deadlines [33]. And, the recent industrial effort known as Datacenter Bridging extends Ethernet to support traffic from other protocols that have different link layer assumptions [2]. All of these approaches focus on single-path mechanisms that are bound by the performance of flow hashing.

Datacenter protocols focused on spreading load across multiple paths have been proposed. Hedera performs periodic flow remapping of elephant flows [10]. MPTCP takes a step further, making TCP aware of multiple paths [29]. While these approaches provide multipath support, they operate at timescales that are too coarse-grained to improve the short flow completion time tail.

8.3 HPC Interconnects

DeTail borrows some ideas from HPC interconnects. Credit-based flow control has been extensively studied and is often deployed to create lossless fabrics [8]. Adaptive load balancing algorithms such as UGAL and PAR have also been proposed [8]. To the best of our knowledge, these mechanisms have not been evaluated for web-facing datacenter networks focused on reducing the flow completion tail.

A commodity HPC interconnect, Infiniband, has made its way into datacenter networks [5]. While Infiniband provides a priority-aware lossless interconnect, it does not perform Adaptive Load Balancing (ALB). Without ALB, hotspots can occur, leading a subset of flows to hit the long tail. Host-based approaches to performing load-balancing, such as [32] have been proposed. But these approaches are limited because they are not sufficiently agile.

9. CONCLUSION

In this paper, we presented DeTail, an approach for reducing the tail of completion times of the short, latency-sensitive flows critical for page creation. DeTail employs cross-layer, in-network mechanisms to reduce packet losses and retransmissions, prioritize latency-sensitive flows, and evenly balance traffic across multiple paths. By making its flow completion statistics robust to congestion, DeTail can reduce 99.9^{th} percentile flow completion times by over 50% for many workloads.

DeTail's approach will likely achieve significant improvements in the tail of flow completion times for the foreseeable future. Increases in network bandwidth are unlikely to be sufficient. Buffers will drain faster, but they will also fill up more quickly, ultimately causing the packet losses and retransmissions that lead to long tails. Prioritization will continue to be important as background flows will likely remain the dominant fraction of traffic. And load imbalances due to topological asymmetries will continue to create hotspots. By addressing these issues, DeTail enables web sites to deliver richer content while still meeting interactivity deadlines.

10. ACKNOWLEDGEMENTS

This work is supported by MuSyC: "Multi-Scale Systems Center", MARCO, Award #2009-BT-2052 and AmpLab: "AMPLab: Scalable Hybrid Data Systems Integrating Algorithms, Machines and People", DARPA, Award #031362. We thank Ganesh Ananthanarayanan, David Culler, Jon Kuroda, Sylvia Ratnasamy, Scott Shenker, and our shepherd Jon Crowcroft for their insightful comments and suggestions. We also thank Mohammad Alizadeh and David Maltz for helping us understand the DCTCP workloads.

11. REFERENCES

[1] Cisco nexus 5000 series architecture. http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9670/white_paper_c11-462176.html.

[2] Data center bridging. http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns783/at_a_glance_c45-460907.pdf.

[3] Datacenter networks are in my way. http://mvdirona.com/jrh/TalksAndPapers/JamesHamilton_CleanSlateCTO2009.pdf.

[4] Fulcrum focalpoint 6000 series. http://www.fulcrummicro.com/product_library/ FM6000_Product_Brief.pdf.

[5] Infiniband architecture specification release 1.2.1. http://infinibandta.org/.

[6] Ns3. http://www.nsnam.org/.

[7] Priority flow control: Build reliable layer 2 infrastructure. http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9670/white_paper_c11-542809.pdf.

[8] ABTS, D., AND KIM, J. High performance datacenter networks: Architectures, algorithms, and opportunities. *Synthesis Lectures on Computer Architecture 6*, 1 (2011).

[9] AL-FARES, M., LOUKISSAS, A., AND VAHDAT, A. A scalable, commodity data center network architecture. In *SIGCOMM* (2008).

[10] AL-FARES, M., RADHAKRISHNAN, S., RAGHAVAN, B., HUANG, N., AND VAHDAT, A. Hedera: Dynamic flow scheduling for data center networks. In *NSDI* (2010).

[11] ALIZADEH, M. Personal communication, 2012.

[12] ALIZADEH, M., GREENBERG, A., MALTZ, D. A., PADHYE, J., PATEL, P., PRABHAKAR, B., SENGUPTA, S., AND SRIDHARAN, M. Data center tcp (dctcp). In *SIGCOMM* (2010).

[13] ALIZADEH, M., KABBANI, A., EDSALL, T., PRABHAKAR, B., VAHDAT, A., AND YASUDA, M. Less is more: Trading a little bandwidth for ultra-low latency in the data center. In *NSDI* (2012).

[14] BENZEL, T., BRADEN, R., KIM, D., NEUMAN, C., JOSEPH, A., SKLOWER, K., OSTRENGA, R., AND SCHWAB, S. Experience with deter: a testbed for security research. In *TRIDENTCOM* (2006).

[15] BRAKMO, L. S., O'MALLEY, S. W., AND PETERSON, L. L. Tcp vegas: new techniques for congestion detection and avoidance. In *SIGCOMM* (1994).

[16] CHEN, Y., GRIFFITH, R., LIU, J., KATZ, R. H., AND JOSEPH, A. D. Understanding tcp incast throughput collapse in datacenter networks. In *WREN* (2009).

[17] CLARK, D. The design philosophy of the darpa internet protocols. In *SIGCOMM* (1988).

[18] DEAN, J. Software engineering advice from building large-scale distributed systems. http://research.google.com/people/jeff/stanford-295-talk.pdf.

[19] DEMERS, A., KESHAV, S., AND SHENKER, S. Analysis and simulation of a fair queueing algorithm. In *SIGCOMM* (1989).

[20] FLOYD, S., AND HENDERSON, T. The newreno modification to tcp's fast recovery algorithm, 1999.

[21] FLOYD, S., AND JACOBSON, V. Random early detection gateways for congestion avoidance. *IEEE/ACM Trans. Netw. 1* (August 1993).

[22] GREENBERG, A., HAMILTON, J. R., JAIN, N., KANDULA, S., KIM, C., LAHIRI, P., MALTZ, D. A., PATEL, P., AND SENGUPTA, S. Vl2: a scalable and flexible data center network. In *SIGCOMM* (2009).

[23] GUO, C., LU, G., LI, D., WU, H., ZHANG, X., SHI, Y., TIAN, C., ZHANG, Y., AND LU, S. Bcube: A high performance, server-centric network architecture for modular data centers. In *SIGCOMM* (2009).

[24] GUO, C., WU, H., TAN, K., SHI, L., ZHANG, Y., AND LU, S. Dcell: a scalable and fault-tolerant network structure for data centers. In *SIGCOMM* (2008).

[25] JACOBSON, V., AND BRADEN, R. T. Tcp extensions for long-delay paths, 1988.

[26] KOHAVI, R., AND LONGBOTHAM, R. Online experiments: Lessons learned, September 2007. http://exp-platform.com/Documents/IEEEComputer2007 OnlineExperiments.pdf.

[27] KOHLER, E., MORRIS, R., CHEN, B., JANNOTTI, J., AND KAASHOEK, M. F. The click modular router. *ACM Trans. Comput. Syst. 18* (August 2000).

[28] MCKEOWN, N. White paper: A fast switched backplane for a gigabit switched router. http://www-2.cs.cmu.edu/ srini/15-744/readings/McK97.pdf.

[29] RAICIU, C., BARRE, S., PLUNTKE, C., GREENHALGH, A., WISCHIK, D., AND HANDLEY, M. Improving datacenter performance and robustness with multipath tcp. In *SIGCOMM* (2011).

[30] SALTZER, J. H., REED, D. P., AND CLARK, D. D. End-to-end arguments in system design. *ACM Trans. Comput. Syst. 2* (November 1984).

[31] VASUDEVAN, V., PHANISHAYEE, A., SHAH, H., KREVAT, E., ANDERSEN, D. G., GANGER, G. R., GIBSON, G. A., AND MUELLER, B. Safe and effective fine-grained TCP retransmissions for datacenter communication. In *SIGCOMM* (2009).

[32] VISHNU, A., KOOP, M., MOODY, A., MAMIDALA, A. R., NARRAVULA, S., AND PANDA, D. K. Hot-spot avoidance with multi-pathing over infiniband: An mpi perspective. In *CCGRID* (2007).

[33] WILSON, C., BALLANI, H., KARAGIANNIS, T., AND ROWTRON, A. Better never than late: meeting deadlines in datacenter networks. In *SIGCOMM* (2011).